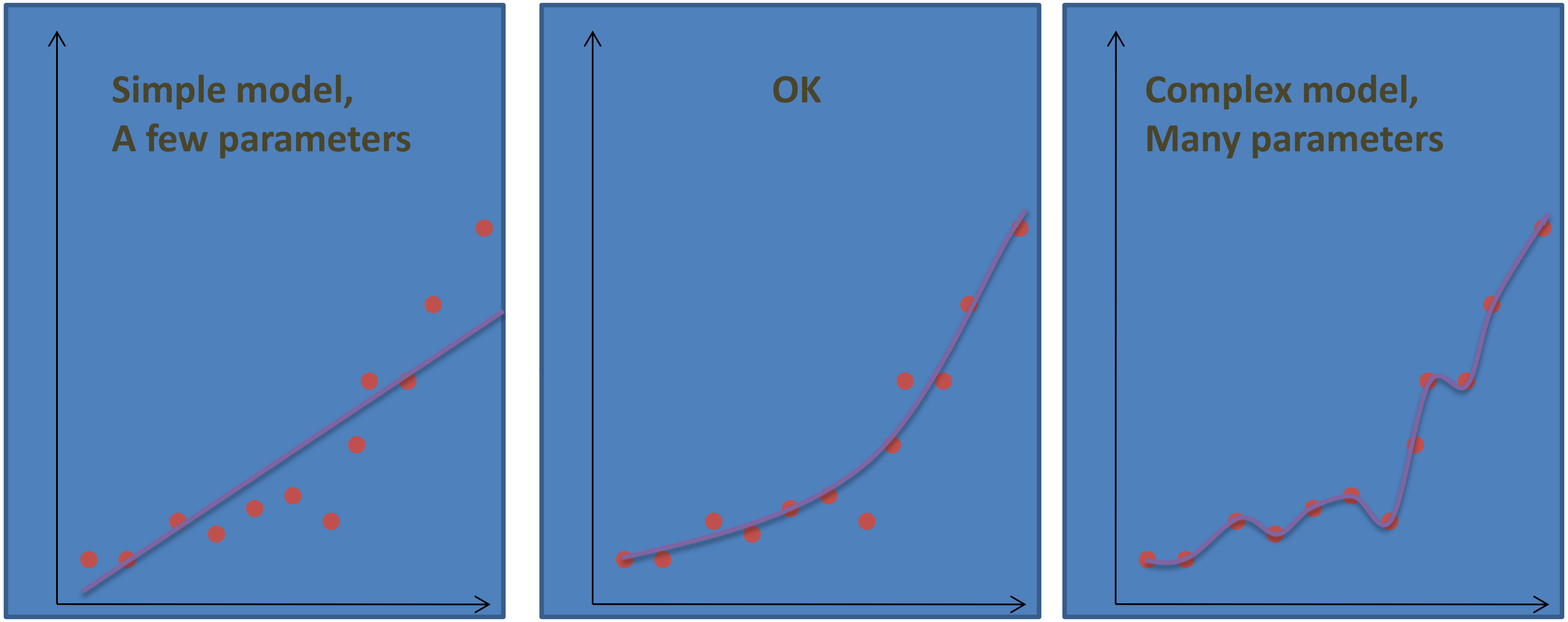
Lecture 5:
Cross Validation, Jackknife, bootstrap

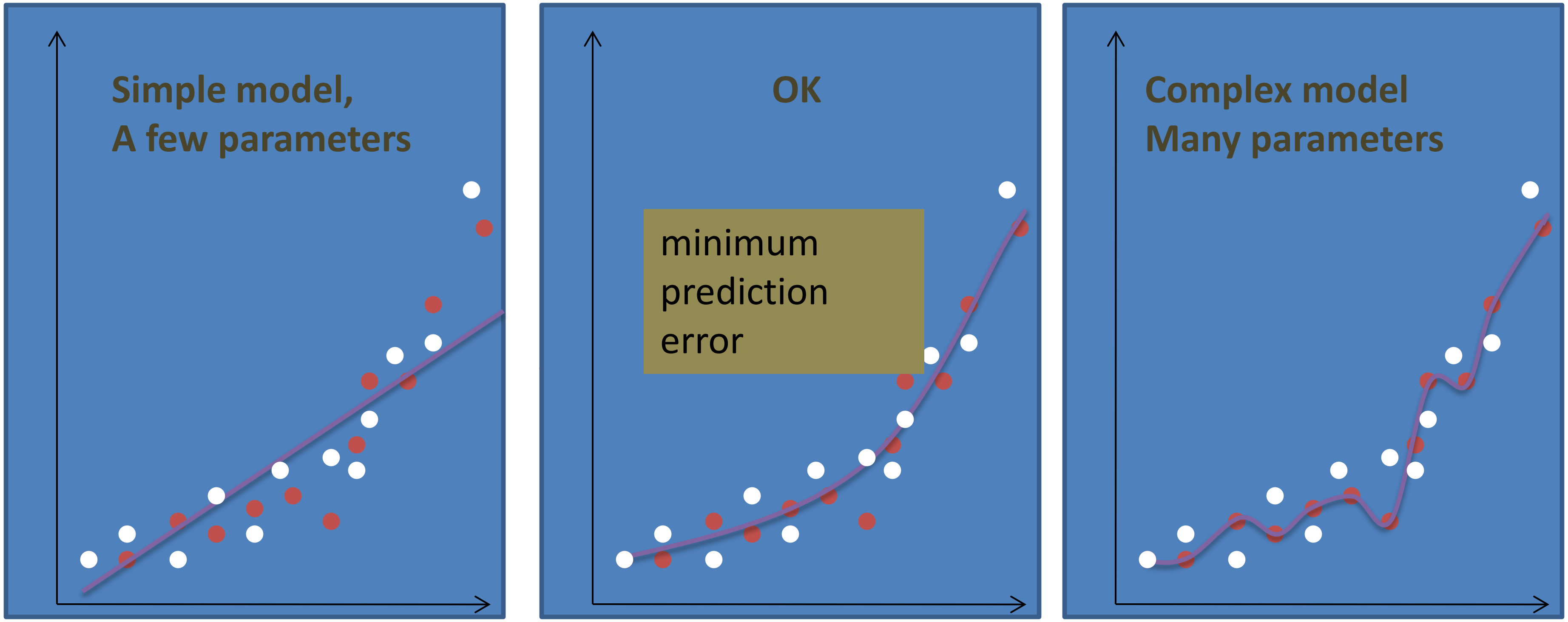
Model selection

Overfitting



What is a good model

- $Y = f(X) + \varepsilon$



Model selection

WHEN THE DATA SET IS BIG ENOUGH

Divide into training, validation and test

Training	Validation	Test

- Recommended proportions: 1/3, 1/3, 1/3 or 60%,20%,20%
- Alternative- obtain datasets by sampling from the original set

Cross-Validation

- What to do when the data set is not enough big?
- **Idea**: estimate prediction error at one or several points by fitting the model to the remaining data

Cross-Validation

K-fold cross-validation (rough scheme, picture follows):

1. Divide data-set in K roughy equal-sized subsets
2. Remove subset #i and fit the model using remaining data.
3. Predict the function values for subset #i using constructed model.
4. Repeat steps 2-3 for different i
5. CV= squared difference between observed values and predicted values (another function is possible)

Cross-Validation

1	2	3	4	5
Train	Train	Test	Train	Train

<u>Note</u>: if K=N then method is *leave-one-out* cross-validation.

K-fold cross-validation: $\kappa : \{1, \ldots, N\} \mapsto \{1, \ldots, K\}$

$$\mathrm{CV} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

Cross-Validation

How to use CV score for model selection?

Having model depending on tuning (complexity) parameter, choose the one with smallest CV:

$$\mathrm{CV}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i, \alpha))$$

Cross-Validation

Issues

- Limitations: model should be globally defined
- How to define best K?

Jackknife methods

- Idea: similar to CV, but used in statistical inference
  - Bias estimation
  - Variance estimation

"Jackknife methods make use of systematic partitions of a dataset to estimate properties of an estimator computed from the full sample"

- Suppose, we are given a random sample **Y**=(Y_1,…,Y_n) and some estimator T(**Y**)

Jackknife methods

First-order jackknife

1. Obtain $\mathbf{Y}_{(-j)}$ by dropping group of observations j from $\mathbf{Y}$
2. For each j, compute $T_{(-j)} = T(\mathbf{Y}_{(-j)})$
3. Compute pseudovalues and J(T), called *jackknifed* T:

$$\overline{T}_{(\bullet)} = \frac{1}{r}\sum_{j=1}^{r} T_{(-j)} \qquad T_j^* = rT - (r-1)T_{(-j)}$$

$$\mathrm{J}(T) = \frac{1}{r}\sum_{j=1}^{r} T_j^* = \overline{T}^*$$

- Equivalently, $\mathrm{J}(T) = rT - (r-1)\overline{T}_{(\bullet)}$

Jackknife variance estimate

- We can use $T_{(-j)}$ or pseudovalues as estimates of T for different samples (both give equivalent expression).

- Variance becomes

$$\widehat{\mathrm{V}(T)}_{\mathrm{J}} = \frac{\sum_{j=1}^{r}\left(T_j^* - \mathrm{J}(T)\right)^2}{r(r-1)}$$

Sometimes, one takes $\dfrac{\sum_{j=1}^{r}(T_j^* - T)^2}{r(r-1)}$

!The variance is often overestimated

Jackknife bias correction

First-order jackknife

- The bias reduced to order n^{-1} (we take r=n)

$$\mathrm{Bias}(T) = \mathrm{E}(T) - \theta \quad = \sum_{q=1}^{\infty} \frac{a_q}{n^q}$$

$$\begin{aligned}
\mathrm{Bias}(\mathrm{J}(T)) &= \mathrm{E}(\mathrm{J}(T)) - \theta \\
&= n(\mathrm{E}(T) - \theta) - \frac{n-1}{n} \sum_{j=1}^{n} \mathrm{E}(T_{(-j)} - \theta) \\
&= n \sum_{q=1}^{\infty} \frac{a_q}{n^q} - (n-1) \left(\sum_{q=1}^{\infty} \frac{a_q}{(n-1)^q} \right) \\
&= a_2 \left(\frac{1}{n} - \frac{1}{n-1} \right) + a_3 \left(\frac{1}{n^2} - \frac{1}{(n-1)^2} \right) + \ldots \\
&= -a_2 \left(\frac{1}{n(n-1)} \right) + a_3 \left(\frac{1}{n^2} - \frac{1}{(n-1)^2} \right) + \ldots
\end{aligned}$$

Jackknife estimation of bias

- We see that

$$\mathrm{E}(\mathrm{J}(T)) - \theta \; = \; \mathrm{E}(T) - \theta + (n-1)\left(\mathrm{E}(T) - \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}(T_{(-j)})\right)$$

- Hence, bias is $B_{\mathrm{J}} = (n-1)\left(\overline{T}_{(\bullet)} - T\right)$

Higher-order jackknife

The order of the bias can be further reduced

- Second-order jackknife

$$\mathrm{J}^2(T) = \frac{n^2 \mathrm{J}(T) - (n-1)^2 \sum_{j=1}^{n} \mathrm{J}(T)_{(-j)}/n}{n^2 - (n-1)^2}$$

- Higer order jackkifes –combining jackknifes of lower orders:

$$T_w = \frac{T_1 - wT_2}{1 - w}$$

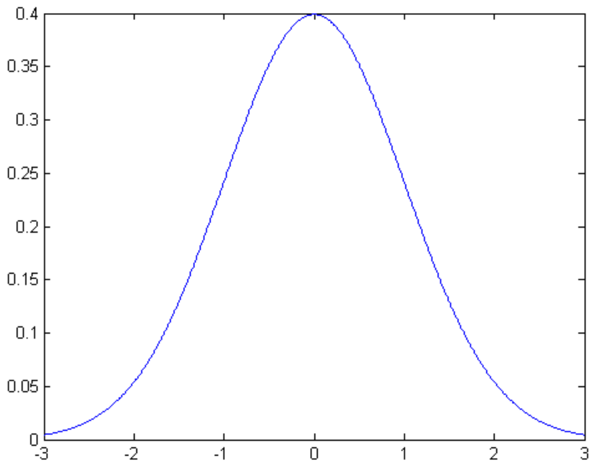
Higher-order jackknife

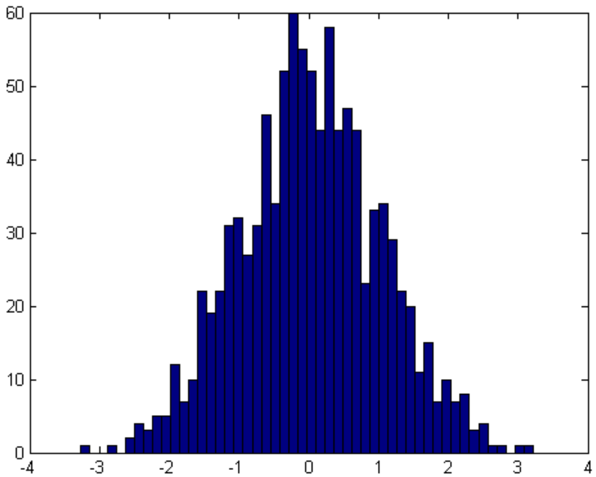
Comments

- High order jackknifes reduce the bias but they increase the variance

- Delete-1 jackknife is not always appropriate (median). Use delete-k

Bootstrap





Unknown: Distribution of **X** (CDF P)

We have: Data set **D** (ECDF P_n)

They are similar!

Bootstrap

What do we want:

- Determine the distribution of functional of *P*, $\Theta(P)$
  - *Bias*
  - *Variance*

Why do we need *bootstrap*?

- Often the true distribution is unknown, therefore distribution of $\Theta(P)$ is unknown
- Even if the distribution is known, but $\Theta(P)$ has complex structure, no problem for bootstrap

Bootstrap

What do we do:

- Unknown $\theta = \Theta(P) \quad = \int g(y)\,\mathrm{d}P(y)$

- We approximate $T \;= \Theta(P_n) \;= \int g(y)\,\mathrm{d}P_n(y)$

(we use the sample to compute θ)

Algorithm (nonparametric bootstrap):

1. Using observation set **D=**(X_1,…X_n), sample with replacement and get bootstrap sample $\mathbf{D_1}$ **=**(X^*_1,…X^*_n),
2. Repeat step 1 B times
3. The distribution of θ is given by T($\mathbf{D_1}$)… T($\mathbf{D_B}$)

Bootstrap

<u>Example</u>:

Having sample $X_1,\dots.X_n$ from normal distribution

- Our estimator of EX is T=mean(X). How to find the variance of T?
  - *Analytically*
  - *Using bootstrap*
- How to find the variance of log(mean(T))?
- What happens if you don’t know the distribution of T?

Bootstrap

Algorithm (parametric bootstrap):

Assume that you know that X comes from F(α) but α is unknown

1. Estimate α from data **D=**$(X_1,\ldots X_n)$,by maximum likelihood
2. Generate $\mathbf{D_1}$ **=**$(X^*_1,\ldots X^*_n)$ by sampling from F(α)
3. Repeat step 2 B times
4. The distribution of θ is given by $T(\mathbf{D_1})\ldots T(\mathbf{D_B})$

Note: In regression, there is a *semiparametric* bootstrap (residual resampling)

Bootstrap: regression context

- Model $Y = f(x) + \epsilon, \epsilon \sim F(\alpha)$
- Data D $= \{(Y_i, X_i), i = 1, \dots, n\}$
- Idea: for produce several bootstrap sets that are similar to D

Algorithm (nonparametric bootstrap): F is unknown

1. Using observation set **D**, sample **pairs** (X_i, Y_i) with replacement and get bootstrap sample $\mathbf{D_1}$
2. Repeat step 1 B times

Bootstrap: regression context

Algorithm (parametric bootstrap): F is known

1. Fit a model to D → get $\hat{f}(X_i)$. Estimate α from data **D** by maximum likelihood
2. Set $X_i^* = X_i$, generate $\epsilon_i \sim F(\alpha)$ and compute $Y_i^* = \hat{f}(X_i) + \epsilon_i$.
3. $D_i = \{(X_i^*, Y_i^*), i = 1, \dots, n\}$
4. Repeat step 2 B times

Bootstrap: regression context

Algorithm (semiparametric bootstrap): F is unknown

1. Fit a model to D → get $\hat{f}(X_i)$. Estimate residuals $R = \{r_i, i = 1, \ldots, n\}$
2. Set $X_i^* = X_i$, sample from R and get ϵ_i and compute $Y_i^* = \hat{f}(X_i) + \epsilon_i$.
3. $D_i = \{(X_i^*, Y_i^*), i = 1, \ldots, n\}$
4. Repeat step 2 B times

Bootstrap bias corrections

- Theory shows

$$T_1 = 2T(P_n) - \mathrm{E}\Big(T\big(P_n^{(1)}\big) \,\Big|\, P_n\Big)$$

- The last term is computed by
  1. Using observation set **D=**(X_1,…X_n), sample with replacement and get bootstrap sample **D_1 =**(X^*_1,…X^*_n),
  2. Repeat step 1 B times
  3. Take the mean of T(**D_1**)… T(**D_B**)

- The first term is simply the 2T**(D)**

Bootstrap variance estimation

- Using bootstrap, compute T^{*1}=T($\mathbf{D_1}$)… T^{*m}= T($\mathbf{D_B}$)

$$\widehat{\mathrm{V}}(T) \quad = \frac{1}{m-1}\sum(T^{*j} - \overline{T}^{*})^2$$

Bootstrap confidence intervals

- To estimate 100(1-α) confidence interval for $\Theta(P)$

Bootstrap percentile method

1. Using bootstrap, compute $T^{*1}=T(\mathbf{D_1})\ldots\ T^{*m}=T(\mathbf{D_B})$, reorder them ascending
2. Define A_1=ceil(B* α/2), A_2=floor(B-B* α/2)
3. Confidence interval is given by

$$\left(T^{*A_1}, T^{*A_2}\right)$$

Look at the plot…

Bootstrap confidence intervals

Bootstrap-t method

1. Using bootstrap, compute $T^{*1}=T(\mathbf{D_1})\ldots T^{*m}= T(\mathbf{D_B})$
2. Compute
$$t_j = \frac{T^{*j} - T(\mathbf{D})}{se(T^{*j})}, \; j = 1 \ldots B$$
3. Let A_1 and A_2 be $\alpha/2$ and $1-\alpha/2$ percentiles of t (If B=1000, α=0.1, then A_1 is 50th smallest, A_2 is 950th smallest)
4. Confidence interval is $\left(T(D) - se(T)\cdot t_{A_2}, T(D) - se(T)\cdot t_{A_1}\right)$

Bootstrap confidence intervals

Comments

- *se* is square root of estimated variance
- Estimation $se(T^{*j})$ typically requires second-level bootstrap -> bootstrap-t is computationally intensive
- Bootstrap-t is more accurate than percentile (coverage error)
- Bootstrap BC_a is a more advanced bootstrap CI method

Recommended reading

- Chapters 12 and 13
- R: package ”boot”